Privacy at Scale: Local Differential Privacy in Practice

Graham Cormode
University Of Warwick
G.Cormode@warwick.ac.uk

Tejas Kulkarni
University Of Warwick
T.Kulkarni.2@warwick.ac.uk

Divesh Srivastava
AT&T Labs-Research
divesh@research.att.com

ABSTRACT

The most impactful data science often relies on analysing data from individuals that is considered highly sensitive — medical history, personal interests and preferences, and opinions. In many cases it is not feasible to gather the necessary sensitive information without providing strong guarantees of privacy to the users in question. The model of differential privacy can provide such guarantees, and most recently the topic of local differential privacy (LDP) — where users randomly perturb their inputs to provide plausible deniability of their data without the need for a trusted party — has come to the fore.

Local differential privacy has been adopted by several major technology organizations, so the technology is used by hundreds of millions of users daily. These companies include Google through their RAPPOR system, to collect web browsing behaviour [12]; Apple's implementation, that allows Apple and app developers to collect usage and typing history [1]; and Microsoft's collection of a variety of telemetry data over time [10].

The aims of this tutorial are to introduce the key technical underpinnings of these deployed systems, and provide intuition on how they work; to survey current research that addresses related problems within the LDP model; and to identify open problems and research directions for privacy. For this tutorial, we use the deployed systems to exemplify and motivate the ideas that derive from algorithms and theory. Participants will learn how an idea from fifty years ago has found application in the 21st Century, and how major companies are scaling this up to Internet scale.

KEYWORDS

Data collection, privacy, differential privacy, local differential privacy

ACM Reference format:
Graham Cormode, Tejas Kulkarni, and Divesh Srivastava. 2018. Privacy at Scale: Local Differential Privacy in Practice. In *Proceedings of Submission to ACM SIGMOD, Houston, Texas USA, June 2018 (Submitted to SIGMOD'18),* 3 pages.
DOI: 10.1145/nnnnnnn.nnnnnnn

Submitted to SIGMOD'18, Houston, Texas USA
 978-x-xxxx-xxxx-x/YY/MM…$15.00
DOI: 10.1145/nnnnnnn.nnnnnnn

1 OUTLINE OF THE TUTORIAL

We propose a short (1.5 hour) tutorial to introduce SIGMOD participants to this new but rapidly developing topic. Our approach is practice-led, inspired by the large-scale deployments of Locally Differential Private data collection by major technology companies, including Google, Apple and Microsoft [9, 10, 12]. We will structure the core of the tutorial around three deployed systems, using them to motivate the underlying algorithms, and connecting out to the research literature that underpins them. In more detail, our outline is as follows:

1.1 Introduction and Preliminaries.

We will briefly motivate the need for tools for private data collection and analysis, and introduce the definitions of Differential Privacy, and the special case of Local Differential Privacy (LDP). The first definition equivalent to LDP came from the database community as "amplification" [13], then came to prominence in the work of Duchi et al. [11]. We will introduce the most basic LDP mechanism, randomized response [6, 22], which came from the survey design community, and masks a single bit by tossing a biased coin. We will introduce the mathematical tools to understand LDP, including unbiasedness, variance and confidence tail bounds.

1.2 State of the art deployments.

We will describe three practical realizations of LDP algorithms for collecting popularity statistics, and cover the development of these ideas through the computer science research literature, and subsequent enhancements that have been proposed.

(1) *RAPPOR* from Google, which combines Randomized Response with Bloom Filters to compactly encode massive sets [12]. The application is to identify popular web destinations (URLs), without revealing any individual user's browsing habits. Subsequent work from the same team has described how to efficiently extract the identities of popular destinations without prior knowledge of their URLs [14].
(2) *Apple's DP implementation* was announced in 2016, and is documented in a patent application [1] and subsequent white paper [9]. The technique combines several technical advances: using the Fourier transform to spread out signal information, and sketching techniques to reduce the dimensionality of the massive domain. In parallel, a rich seam of literature has abstracted the problem of *Indentifing Differentially Private Heavy Hitters*, progressively refining and optimizing these techniques [3–5, 19, 21].
(3) *Telemetry collection* from Microsoft, which makes use of histograms and fixed random numbers to collect data over time [10].

1.3 Current Research Directions

We will briefly describe some of the related quite recent results that have been published on applying LDP to other domains.

- *Private location collection.* Data can often be represented as points in multidimensional space—as a simple example, consider user locations in two-dimensional space. Sketching frequencies within multidimensional spaces, allowing rectilinear counting queries to be answered approximately, and identifying "hot spots" are primitives that could then be used to build more sophisticated user activity models. Initial work on this problem has extended LDP private frequency collection [7]. It is open to extend this to build more sophisticated user movement models.
- *Marginal distributions of multidimensional data.* Given users represented as points in multidimensional space, a natural question is to extract distributions over subsets of dimensions. Naively, we could materialize all possible subsets and apply existing approaches, but this rapidly degrades the accuracy. Instead, taking projections of the data via a Fourier basis allows better reconstructions [8].
- *Graph algorithms and synthetic graph modeling.* Much sensitive individual data is best represented as a graph—either a simple graph between users, or a bipartite graph between users and other entities. Recent work has aimed to build accurate graph models under LDP [20].
- *Language modeling.* An application of private data collection is to build better prediction models e.g. for typing on mobile devices. Recent work has shown how to accurately and privately train sophisticated deep neural network models [17].

1.4 Open Problems and New Directions.

We will point to a number of directions for future work, based on emerging trends in the literature.

- *Multiple Rounds.* Most deployed LDP protocols require the user to follow a fixed protocol over their data, and send their (perturbed) response for aggregation. More generally, we could allow multiple rounds of interaction, where the aggregator poses new queries in the light of previous responses. This approach has been proposed for building machine learning models [18]. It is open to understand the power of multiple rounds, compared to what is possible in a single round.
- *Hybrid models.* LDP gives a very strong protection to users, at the expense of lower accuracy compared to a centralized model with a trusted aggregator. Recent work has proposed a hybrid model where some users follow LDP and some users submit to a trusted aggregator, and both sets are "blended" together [2].
- *Theoretical underpinnings.* Several works on LDP have started to appear in the theory literature, addressing questions about the power of LDP [4, 5, 11]: what are the lower bounds on the accuracy guarantees (as a function of privacy parameter and population size); is there any benefit from adding an additive "relaxation" δ to the privacy definition; and minimizing the amount of data collected from each user to a single bit.

1.5 Connection to other models of privacy and security.

Finally, we will briefly connect to other models of privacy and security, that adopt different assumptions and trust models. These include contrasting with the centralized differential privacy model, and achieving privacy by adding centralized noise via encrypted data collection. We will also discuss other approaches from secure multi-party computation, homomorphic encryption and private information retrieval, amongst others, that achieve different tradeoffs.

2 INTENDED AUDIENCE AND BACKGROUND KNOWLEDGE

We intend to make this tutorial accessible to all participants in SIGMOD and PODS. Although there has been a vast amount of research on the topic of privacy, even when narrowing to work on Differential Privacy, the topic of this tutorial is quite accessible, and does not require any familiarity with prior work. The emphasis is on the design of scalable algorithms, with some consideration of how these can be built into robust systems. To appreciate the correctness and accuracy guarantees of the algorithms, some statistical tools are needed. These are at the level of an introductory statistics course: computing the variance of a discrete random variable, and using this to provide confidence bounds. We will give a brief refresher on the necessary tools, and will not provide detailed proofs of the algorithms; rather, we will try to provide our insights into what the guarantees mean, and where the different terms in the guarantees arise from.

3 PRIOR PRESENTATIONS

This tutorial has not been presented previously. There is a small intersection with the tutorial "Differential Privacy in the Wild" (at SIGMOD 2017, VLDB 2016[15, 16]) by Ashwin Machanavajjhala, Xi He and Michael Hay. We estimate this overlap to correspond to at most five minutes of material.

4 BIOGRAPHY OF THE PRESENTERS

Graham Cormode is a Professor in Computer Science at the University of Warwick in the UK, where he work on research topics in data management, privacy and big data analysis. Previously, he was a principal member of technical staff at AT&T Labs-Research. He is a University Liaison Director at the Alan Turing Institute, and in 2017 he was the co-recipient of Adams Prize for Mathematics for his work on Statistical Analysis of Big Data. He has previously presented tutorials on "Streaming in a connected world: Querying and tracking distributed data streams" (with Minos Garofalakis, at VLDB'06 and SIGMOD'07), "Anonymized Data: Generation, Models, Usage" (with Divesh Srivastava, at SIGMOD'09 and ICDE'10), "Sampling for Big Data" (with Nick Duffield, at KDD'14) and "Compact Summaries for Large Datasets", (at PODS'15 and BICOD'15).

Tejas Kulkarni is a PhD student at the University of Warwick, with a focus on Local Differential Privacy. He completed his Masters at Indian Institute of Technology, Madras.

Divesh Srivastava Divesh Srivastava is the head of Database Research at AT&T Labs-Research. He is a Fellow of the ACM, the Vice President of the VLDB Endowment, and the managing editor of the Proceedings of the VLDB Endowment (PVLDB). His research interests and publications span a variety of topics in data management. He received his Ph.D. from the University of Wisconsin, Madison, and his B.Tech. from the Indian Institute of Technology, Bombay. He has presented tutorials including "Data Stream Query Processing" (with Nick Koudas) at VLDB 2003 and ICDE 2005, "Record Linkage: Similarity Measures and Algorithms" (with Nick Koudas and Sunita Sarawagi) at VLDB 2005 and SIGMOD 2006, "Anonymized Data: Generation, Models, Usage" (with Graham Cormode), at SIGMOD 2009 and ICDE 2010, and "Big Data Integration" (with Xin Luna Dong) at ICDE 2013 and VLDB 2013.

REFERENCES

[1] A Thakurta, A Vyrros, U Vaishampayan, G Kapoor, J Freudiger, V Rangarajan Sridhar, D Davidson. Private dictionary population satisfying local differential privacy, March 2017. US Patent 9,594,741 B1.

[2] B. Avent, A. Korolova, D. Zeber, T. Hovden, and B. Livshits. BLENDER: enabling local search with a hybrid differential privacy model. In *26th USENIX Security Symposium, USENIX Security 2017, Vancouver, BC, Canada, August 16-18, 2017.*, pages 747–764, 2017.

[3] R. Bassily, K. Nissim, U. Stemmer, and A. G. Thakurta. Practical locally private heavy hitters. In *NIPS*, pages 2285–2293, 2017.

[4] R. Bassily and A. Smith. Local, private, efficient protocols for succinct histograms. In *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*, pages 127–135. ACM, 2015.

[5] M. Bun, J. Nelson, and U. Stemmer. Heavy hitters and the structure of local privacy. *CoRR*, abs/1711.04740, 2017.

[6] A. Chaudhuri and R. Mukerjee. *Randomized response: Theory and techniques*. Marcel Dekker, 1988.

[7] R. Chen, H. Li, A. K. Qin, S. P. Kasiviswanathan, and H. Jin. Private spatial data aggregation in the local setting. In *32nd IEEE International Conference on Data Engineering, ICDE 2016, Helsinki, Finland, May 16-20, 2016*, pages 289–300, 2016.

[8] G. Cormode, T. Kulkarni, and D. Srivastava. Marginal release under local differential privacy. *CoRR*, abs/1711.02952, 2017.

[9] Differential Privacy Team, Apple. Learning with privacy at scale. 2017.

[10] B. Ding, J. Kulkarni, and S. Yekhanin. Collecting telemetry data privately. In *Advances in Neural Information Processing Systems 30*, December 2017.

[11] J. C. Duchi, M. I. Jordan, and M. J. Wainwright. Local privacy and statistical minimax rates. In *FOCS*. IEEE, 2013.

[12] U. Erlingsson, V. Pihur, and A. Korolova. Rappor: Randomized aggregatable privacy-preserving ordinal response. In *ACM CCS*, 2014.

[13] A. V. Evfimievski, J. Gehrke, and R. Srikant. Limiting privacy breaches in privacy preserving data mining. In *PODS*, pages 211–222. ACM, 2003.

[14] G. Fanti, V. Pihur, and Ú. Erlingsson. Building a rappor with the unknown: Privacy-preserving learning of associations and data dictionaries. *Proceedings on Privacy Enhancing Technologies*, 2016(3):41–61, 2016.

[15] A. Machanavajjhala, X. He, and M. Hay. Differential privacy in the wild: A tutorial on current practices & open challenges. *PVLDB*, 9(13):1611–1614, 2016.

[16] A. Machanavajjhala, X. He, and M. Hay. Differential privacy in the wild: A tutorial on current practices & open challenges. In *SIGMOD Conference*, pages 1727–1730. ACM, 2017.

[17] H. B. McMahan, D. Ramage, K. Talwar, and L. Zhang. Learning differentially private language models without losing accuracy. *CoRR*, abs/1710.06963, 2017.

[18] T. T. Nguyên, X. Xiao, Y. Yang, S. C. Hui, H. Shin, and J. Shin. Collecting and analyzing data from smart device users with local differential privacy. *CoRR*, abs/1606.05053, 2016.

[19] Z. Qin, Y. Yang, T. Yu, I. Khalil, X. Xiao, and K. Ren. Heavy hitter estimation over set-valued data with local differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pages 192–203. ACM, 2016.

[20] Z. Qin, T. Yu, Y. Yang, I. Khalil, X. Xiao, and K. Ren. Generating synthetic decentralized social graphs with local differential privacy. In *CCS*, pages 425–438. ACM, 2017.

[21] T. Wang, J. Blocki, N. Li, and S. Jha. Locally differentially private protocols for frequency estimation. In *26th USENIX Security Symposium, USENIX Security 2017, Vancouver, BC, Canada, August 16-18, 2017.*, pages 729–745, 2017.

[22] S. L. Warner. Randomised response: a survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, 60(309):63–69, Mar. 1965.